Reproducibility Report on Scalable and Efficient Hypothesis Testing with Random Forests

Jerry Gu
cs680
University of Waterloo
jy4gu@uwaterloo.ca

1 Introduction

As part of the ML Reproducibility Challenge, this paper is about replicating the claim from the paper "Scalable and Efficient Hypothesis Testing with Random Forests" (Coleman et al., 2022) that their algorithm involving random forests is computationally efficient and statistically valid, by evaluating and reproducing the experiment corresponding to the paper to verify the validity of the results.

2 Background

There has been many methods developed for assessing variable importance, but they require restricting assumptions. For example, with the most popular method involving the use of the out-of-bag (OOB) error rate (specifically, the difference of the error when a variable is randomized), it was shown to be inaccurate in certain cases when there is multicollinearity. (Coleman et al., 2022)

Meanwhile, a method involving infinite-order U-statistics to attempt to fix this issue has been made. However, such methods are computationally expensive (since it involves estimating the variance, and thus the covariance matrix which costs O(n*n) time). (Coleman et al., 2022)

A permutation test involves testing the null hypothesis that 2 samples are equal in some measure. This is done by first computing the differences in the true measures, then shuffling the samples and computing that difference a certain amount of times to find how many of those have a higher difference than that of the true measures; this is the p-value of the test. Typically, if the p-value of a test is less than 0.05, then we reject the null hypothesis.

Random forests are machine learning algorithms that involve training several decision trees, except each tree is trained on a randomized sample from the training data (known as bootstrap aggregating, or bagging) and a subset of the features. In this way, the trees are approximately independent. Then, predictions are made based on most votes for classification tasks, or averaging the output of the trees for regression tasks.

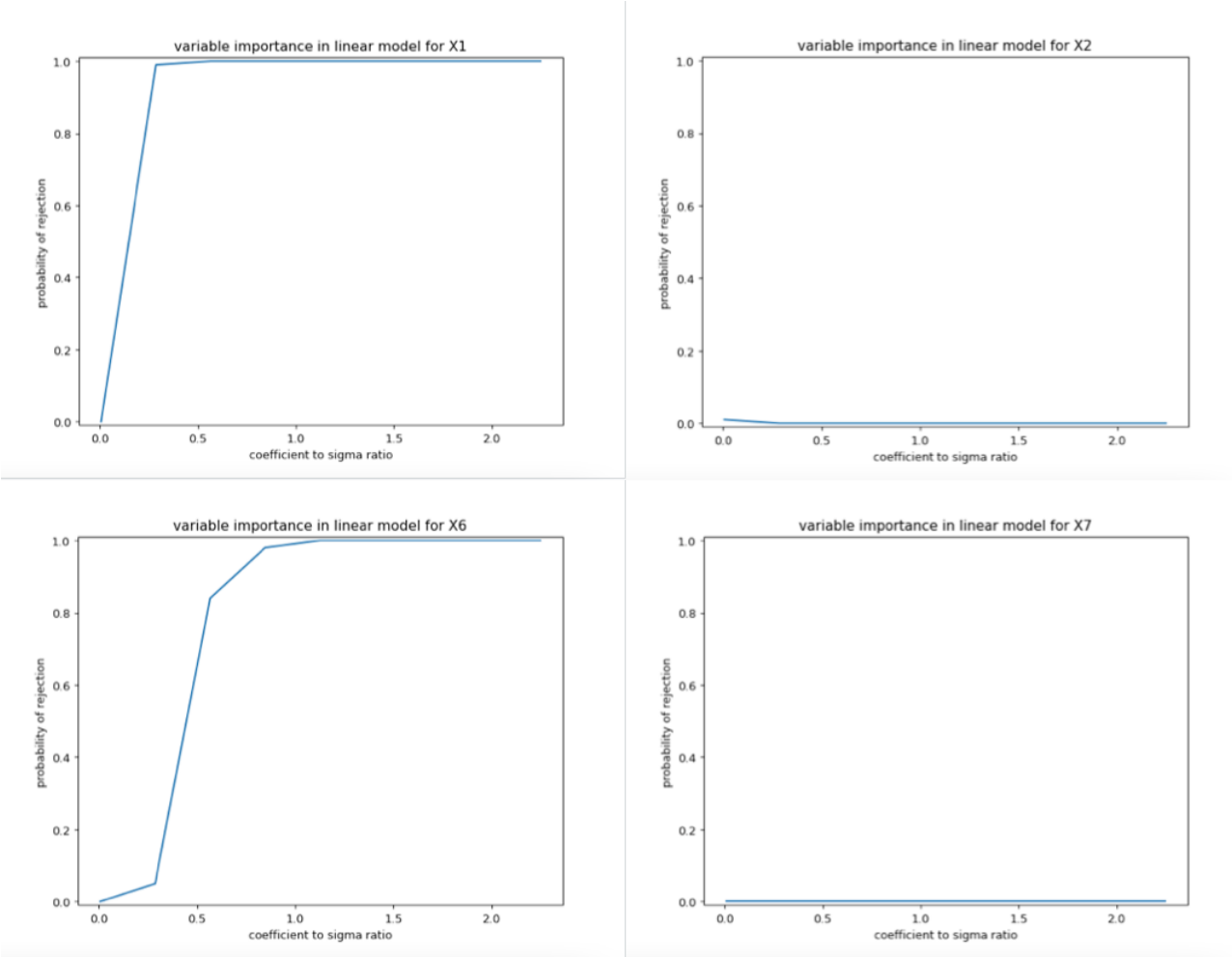
The paper describes an algorithm developed that fixes both issues about restricting assumptions and computation. The algorithm is a permutation test that involves using random forests on 2 datasets: the original, and with columns corresponding to variables of interest randomly shuffled (to destroy any predictive power from those variables) to get B estimates from each dataset; then a permutation test on these 2 sets of estimates are done to determine if they have the same MSE (in which case those variables are not important). (Coleman et al., 2022)

What remains open is to test if the results of the paper are reproducible, which will be done in this report by writing code (in Python) based on the algorithm given in the paper and simulating datasets using the same models as in the paper and determine if the reproduced results are consistent with the paper. Then, OOB error rate will be used to find the variable importances of those models and compare with the results from the algorithm.

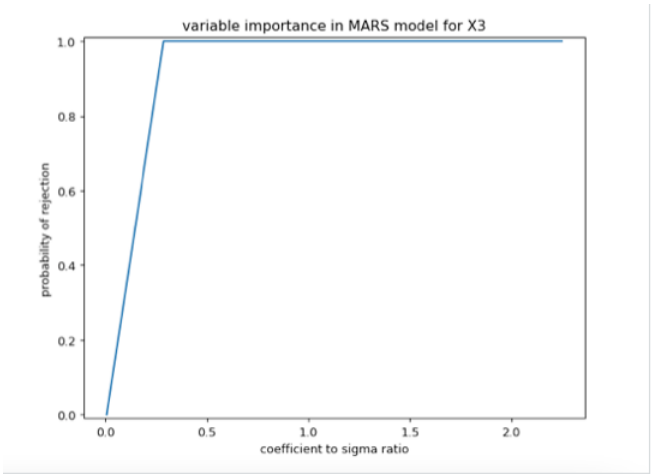
3 Main results

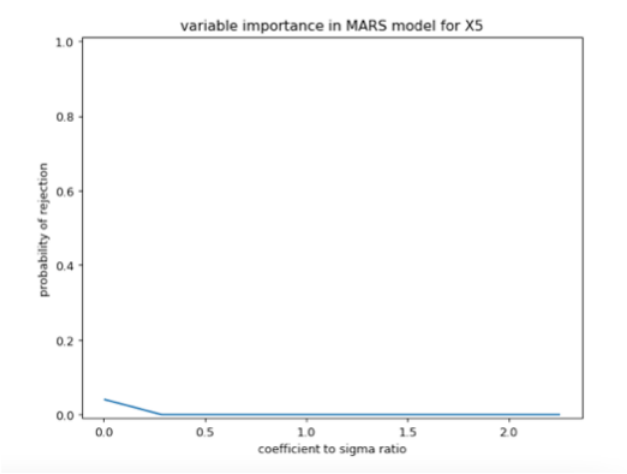
The datasets simulated are from models 1-3 in the simulations section of the paper (Coleman et al., 2022); model 4 was unable to be reproduced and tested since the eBird dataset was not able to be accessed, as it is not publicly available. It is assumed the probabilities of rejecting the null hypothesis that the variable is not important are estimated using 100 trials on the algorithm and taking the proportion of those that have p-value¡0.05. Thus, if the probabilities are above 0.05, then the variable is important.

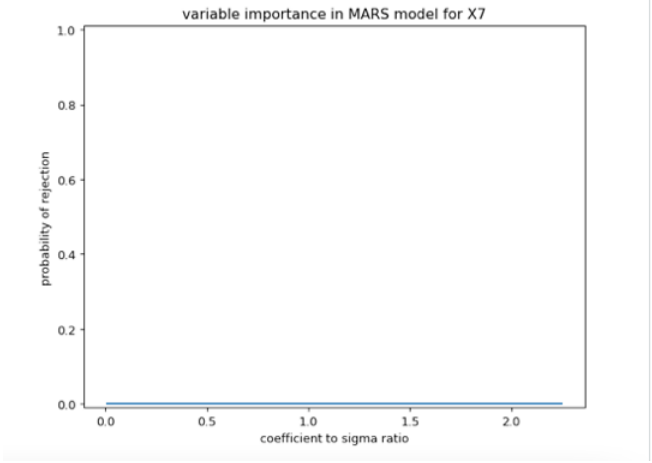
The results for model 1 (linear model) appears to be consistent with the paper; the probability of rejecting the null hypothesis (that the variable is not important) is near 0 for unimportant features (X2 and X7) and approaches 1 (in particular, mostly above 0.05) as ratio of coefficient (i.e. beta) to sigma (i.e. standard deviation of residuals) increases for important features (X1 and X6), like in the paper.

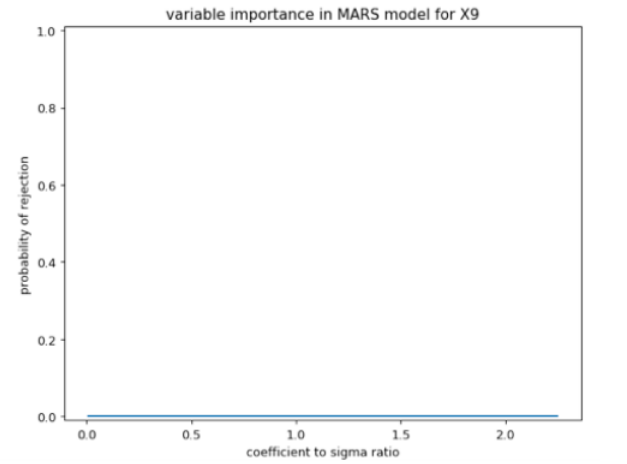


The results for model 2 (MARS model) is mostly consistent with the paper, except with X7, whose probability stayed near 0 instead of approaching 1 like in the paper. The probability for the other important feature, X3, approached 1 (in particular, mostly above 0.05) and the unimportant features (X5, X9) stayed near 0, like in the paper. Thus, even though X7 was important, it was not detected in these reproduced results.

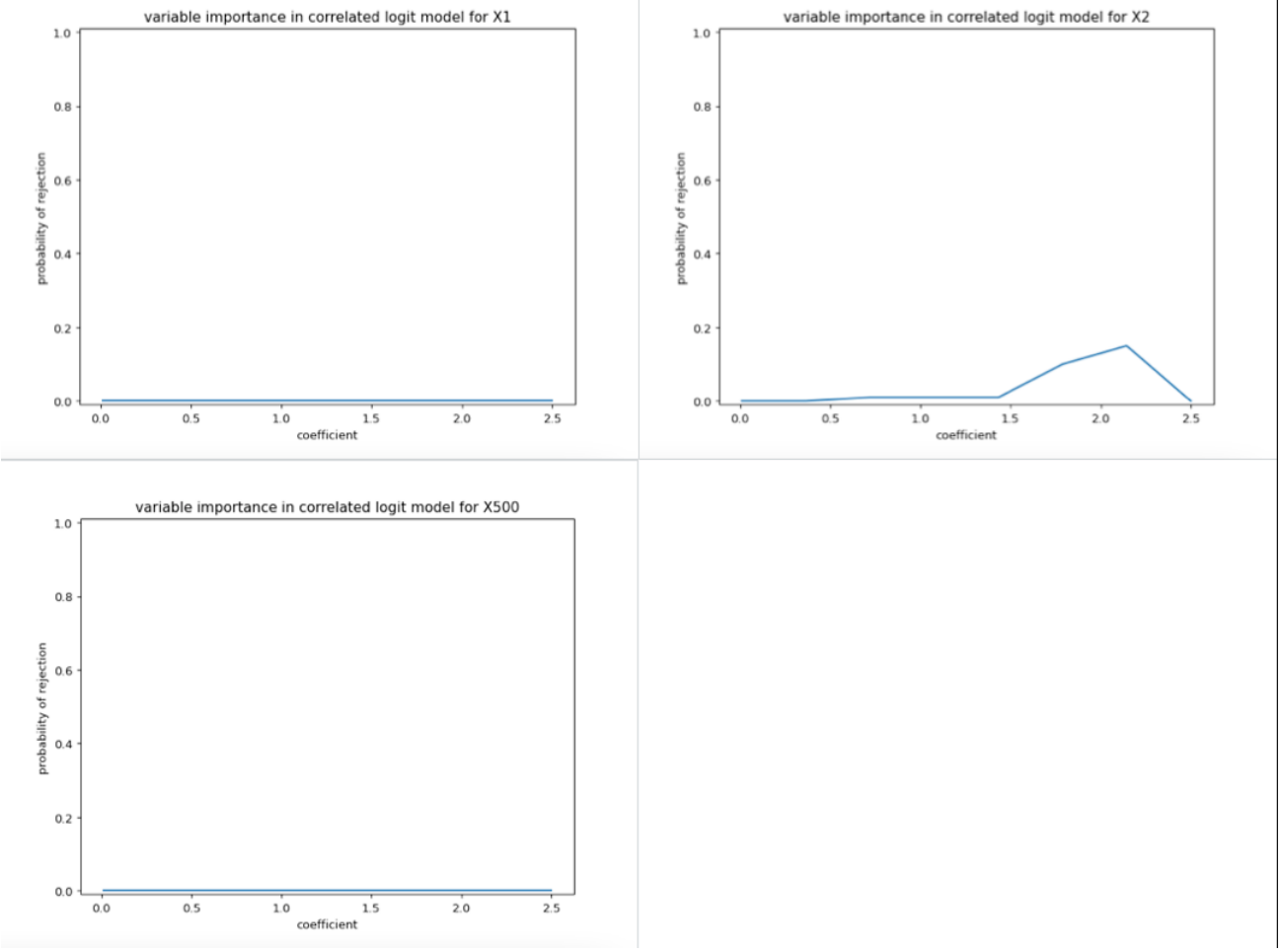






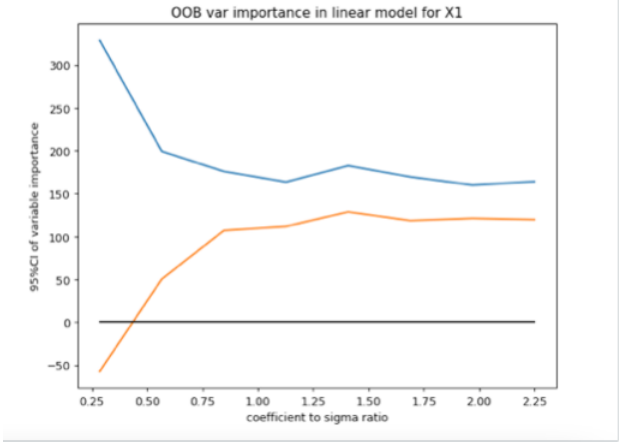


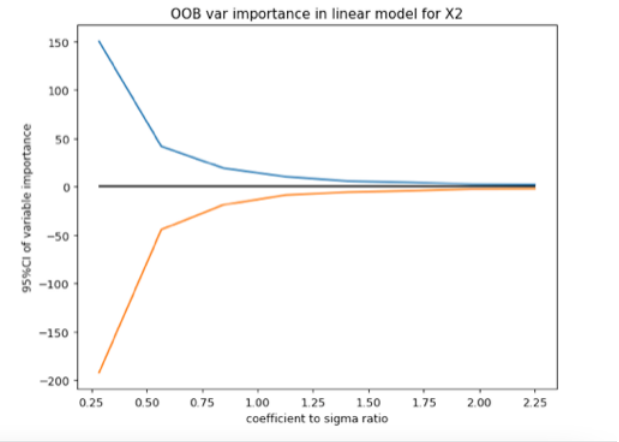
The results for model 3 (correlated logit model) is mostly consistent with the paper, as the probabilities for unimportant features (X1 and X500) stayed near 0 like in the paper, while the important feature (X2) generally increased with the coefficient, although much less compared to that in the paper; particularly, it was above 0.05 for only a few points around the higher coefficients used, so it was sometimes not detected to be important, even though it is important.

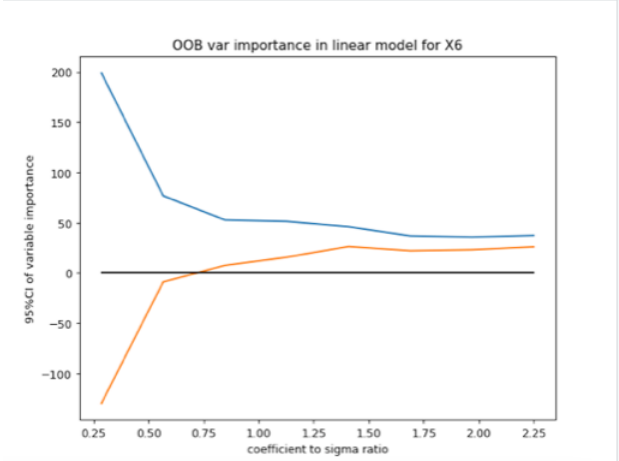


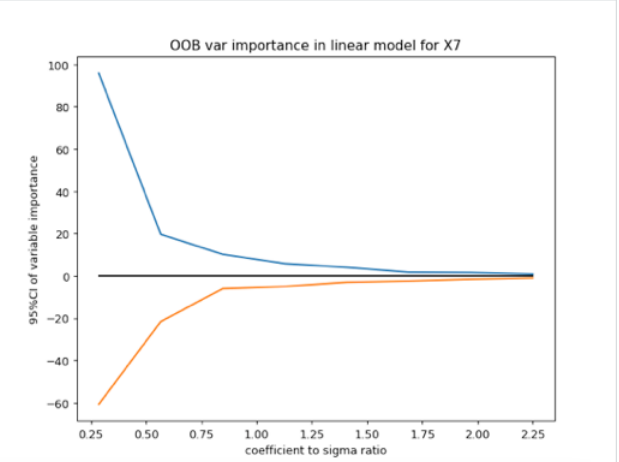
Next, the variable importance in these 3 models will be tested using OOB error rate, to compare with the results using permutation tests as above. These will be visualized with plots of an estimated 95 percent confidence intervals of the variable's importance based on a sample of 100 OOB errors; thus, if the intervals do not include 0, then the variable is important. For the first 2 models, the datasets generated with sigma=0.005 were omitted to make the range of the plots reasonable.

The results were similar to that of the reproduced results above for the first 2 models: important features were found to be important and unimportant features weren't; except X7 in model 2 was again not detected as important, even though it is.

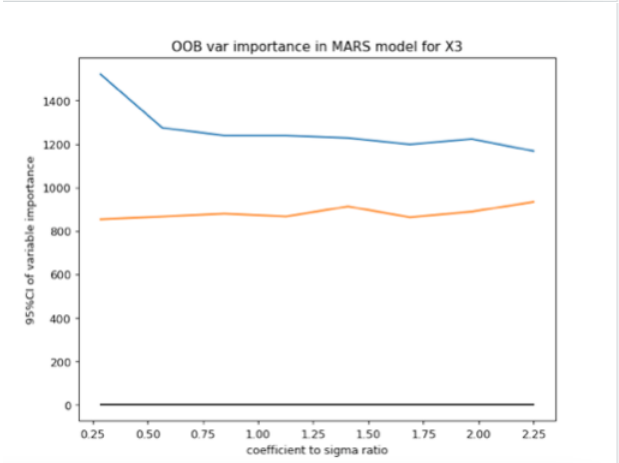

OOB var importance in linear model for X1
95%CI of variable importance
coefficient to sigma ratio

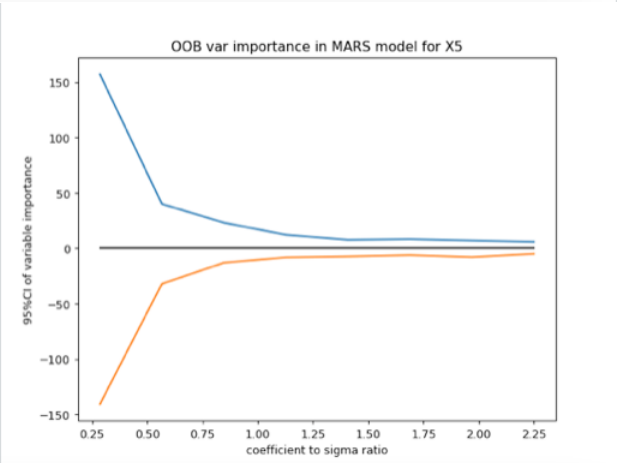

OOB var importance in linear model for X2
95%CI of variable importance
coefficient to sigma ratio

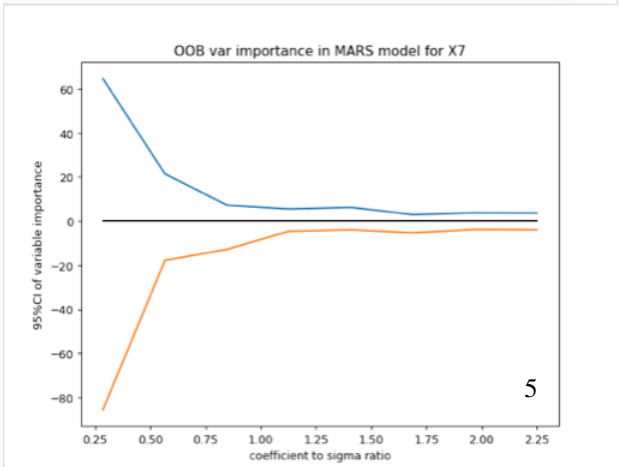

OOB var importance in linear model for X6
95%CI of variable importance
coefficient to sigma ratio

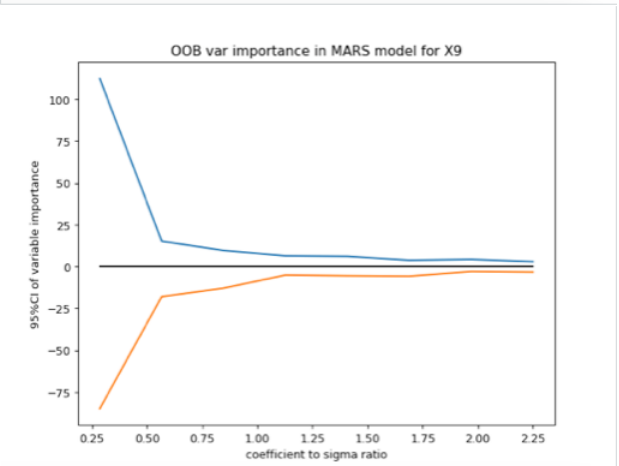

OOB var importance in linear model for X7
95%CI of variable importance
coefficient to sigma ratio

68


OOB var importance in MARS model for X3
95%CI of variable importance
coefficient to sigma ratio

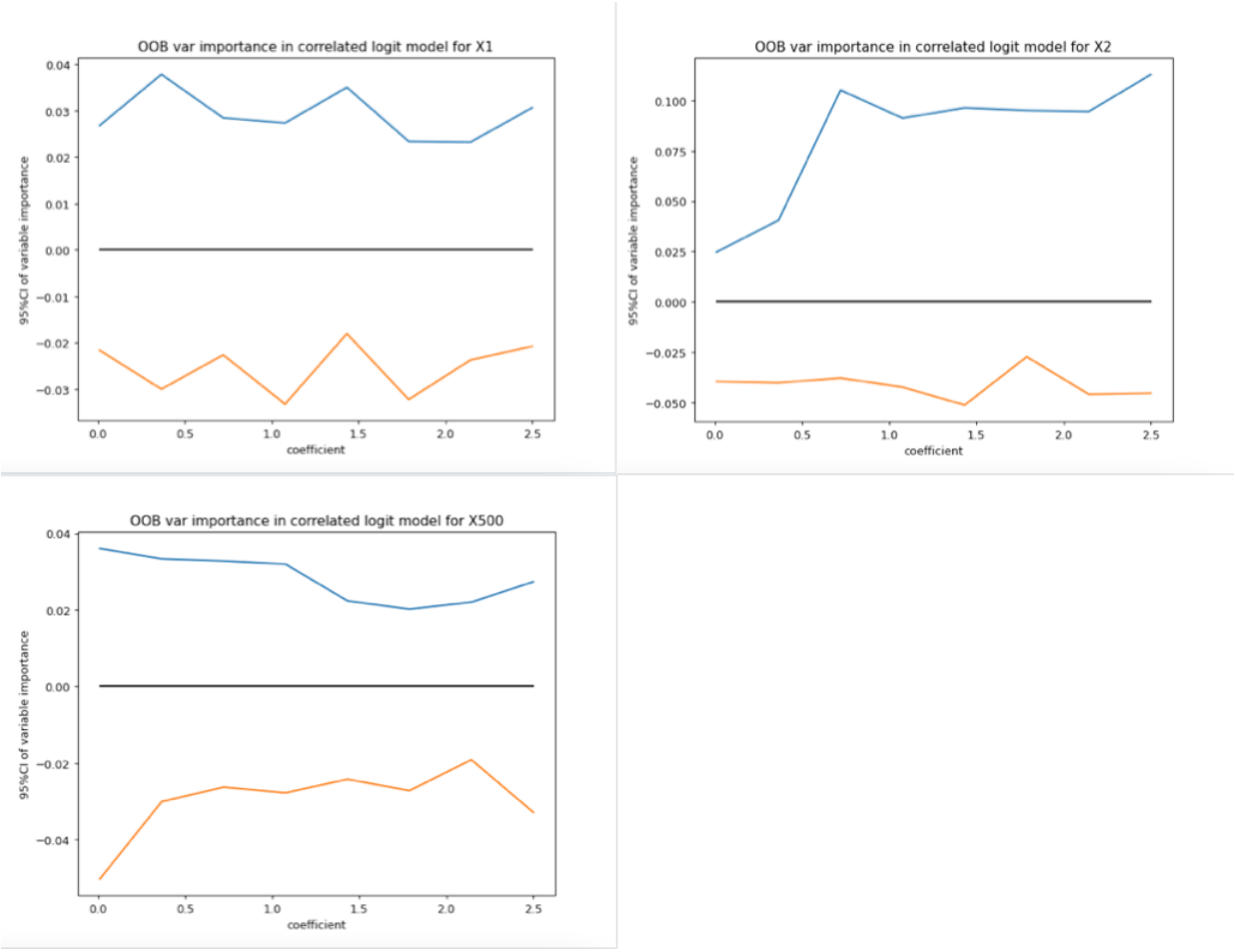

OOB var importance in MARS model for X5
95%CI of variable importance
coefficient to sigma ratio


OOB var importance in MARS model for X7
95%CI of variable importance
coefficient to sigma ratio


OOB var importance in MARS model for X9
95%CI of variable importance
coefficient to sigma ratio

69

For the 3rd model, none of the features were detected as important, even though X2 was actually important.



4 Conclusion

The results were mostly reproducible; however, there were a few discrepancies between the simulations done in the paper "Scalable and Efficient Hypothesis Testing with Random Forests" and the reproduced simulations here; namely, from the proposed algorithm involving permutation tests using random forests, an important feature in a nonlinear regression model was found to be unimportant, and an important feature in a classification model with correlated features was found to only be sometimes important (unlike in the paper where both were found to be important). These results were then compared with the results of finding variable importance using out-of-bag (OOB) error — a method that was shown to be flawed when there is correlation between features. The results were similar, except the OOB error method never found the features in the model with correlated features to be important. Therefore, this experiment has found that the proposed algorithm somewhat fixes the aforementioned flaw with the OOB error method, but may still be unable to detect variable importance in a few cases.

References

Coleman, Tim, et al. "Scalable and Efficient Hypothesis Testing with Random Forests." Edited by Genevra Allen, Journal of Machine Learning Research, June 2022, https://jmlr.org/papers/volume23/20-1060/20-1060.pdf.